On 3D Scene Flow and Structure Estimation*

Ye Zhang and Chandra Kambhamettu
Video / Image Modeling and Synthesis (VIMS) Lab
Department of Computer & Information Sciences
University of Delaware, Newark, Delaware 19716
zhangye/chandra@cis.udel.edu
http://www.cis.udel.edu/∼vims

Abstract

In this paper, novel algorithms computing dense 3D scene flow from multiview image sequences are described. A new hierarchical rule-based stereo matching algorithm is presented to estimate the initial disparity map. Different available constraints under a multiview camera setup are investigated and then utilized in the proposed motion estimation algorithms. We show two different formulations for 3D scene flow computation. One formulation assumes that initial disparity map is accurate while the other does not make this assumption. Image segmentation information is used to maintain the motion and depth discontinuities. Iterative implementations are used to successfully compute 3D scene flow and structure at every point in the reference image. Novel hard constraints are introduced in this paper to make the algorithms more accurate and robust. Promising experimental results are seen by applying our algorithms to real imagery.

1 Introduction

Motion and structure are fundamental problems in computer vision. Most motion estimation methods (*e.g.*, [11, 1, 4, 2]) compute optical flow, *i.e.*, the apparent motion between several frames of an image sequence. In the recent past of visual-motion research, numerous applications including tracking, surveillance, recognition, *etc.*, have been intensively utilizing optical flow information. However, optical flow only provides projected 2D motion information. It is clear that ambiguities exist when dynamic 3D objects/scenes are explained by using 2D optical flow. This is why the counterpart of optical flow in 3D space, 3D scene flow, is introduced [19, 21]. Like optical flow, 3D scene flow is defined at every point in a reference image. The difference is that the velocity vector in scene flow field contains not only x, y, but z velocities. This also means that a multiview camera setup is usually required to compute reliable 3D scene flow.

*Research funding was provided by the National Science Foundation Grants CAREER IRI-9984842 and CISE CDA-9703088.

Over the years, many 3D motion and structure estimation algorithms have been proposed in the literature. Most of them (*e.g.*, [12, 20, 3, 15]) assume that the scene observed is rigid and only rigid motion parameters are estimated. Very little work ([19, 21]) has been done on directly computing the dense 3D motion field from multiview image sequences. Vedula *et al.* [19] designed linear algorithms to compute 3D scene flow under three different scenarios. In their work, optical flow of each view was utilized to estimate scene flow. Then scene structure was estimated from scene flow. By computing optical flow separately, their algorithms still relies on the accuracy of optical flow computation. Also, the linear algorithms may be sensitive to noise. Zhang *et al.* [21] computed 3D scene flow and structure in an integrated manner. In their work, 3D affine motion model was fitted to a local image grid. Then an adaptive global smoothness constraint was applied to the whole image in order to regularize the results. This formulation has problems in occluded areas and at motion/depth boundaries, where the algorithm tends to produce unreliable motion and structure estimations.

Generally speaking, to estimate 3D motion and structure from multiview image sequences, it is desirable to fuse stereo and motion constraints to some extent [20]. However, combining motion/stereo constraints from multiview image sequences requires extra caution. This is because some points in the reference image may be invisible (occluded) in another view. If the algorithm is not aware of this and still combines the motion/stereo constraints from the occluded view, the results could be very wrong.

To deal with the above problems, we propose a method to compute 3D scene flow and structure. The goals of our method include (1) detecting occluded areas in different views, (2) formulating 3D scene flow and structure estimation, and (3) maintaining reliable motion and depth discontinuities.

First, we present a hierarchical rule-based stereo matching algorithm employing image segmentation information to enforce the depth discontinuities. A large

ISBN 0-7695-1272-0/01 $10.00 (C) 2001 IEEE

number of stereo matching algorithms (*e.g.*, [13, 14, 6, 22, 18]) have been proposed in the literature (see [7, 8] for literature surveys on earlier work. An experimental comparison of different stereo algorithms can be found in [17].) However, few stereo algorithms can produce a disparity map, an occlusion map and a confidence map at the same time. Our algorithm is inspired by the work described in [5, 18], where every image segment is assumed to be a planar region with local deformation, and a local model (parametric motion model or planar depth model) is fitted to each segment. Since the fitting is done to individual image segments, discontinuities are normally well maintained. However, planar assumption is not always held in real world. In our algorithm, we try not to make this assumption whenever it is possible. We define a set of rules to adaptively guide the interpolation within each image segment. This set of rules also helps us to find the occluded areas in different views. The output of our algorithm includes a disparity map, an occlusion map and a confidence map.

Then, we formulate 3D scene flow estimation as an energy minimization problem. In fact, this algorithm can be thought of as a natural extension of gradient-based optical flow estimation under a multiview setup. Optical flow constraints from different views are combined together to achieve proper convergence of the scene flow. We show two different formulations for this problem. One formulation assumes that initial depth map computed by using our stereo matching algorithm is accurate while the other does not make this assumption. We also introduce novel hard motion constraints in this algorithm. These hard constraints favor the initial estimation of 2D motion with very high confidence measurement, thus making the algorithm more accurate and robust.

Without loss of generality, some assumptions have been made in our method. First, we assume that all the other cameras are in standard (parallel) set up with the reference camera. Also, all the camera parameters are known and the image sequences captured from different cameras are well rectified. This makes it easy to discuss how to combine constraints from different views. Second, we assume that the motion and depth in each image segment are smooth, thus justifying the enforcement of smoothness constraint inside each image segment. It is worthwhile to note that the smoothness constraint is not unconditionally applied to the entire image. This is the reason why our method can maintain sharp motion and depth boundaries.

The rest of the paper is organized as follows. Section 2 describes the proposed algorithms. Image segmentation is briefly discussed in Section 2.1. Our new stereo matching algorithm is described in Section 2.2. Available motion constraints are investigated in Section 2.3. Hard constraints are explained in Section 2.4. Two different formulations are presented in Section 2.5 and 2.6, respectively. Section 3 reports the experimental results of the proposed method on real imagery. Section 4 concludes the paper.

2 Algorithms

In the following description, we suppose that there are $N(C_0, C_1, ...C_{N-1})$ cameras available and the reference camera is C_0. The image sequences captured by camera k is denoted as $\mathbf{I_{k,t}}(I_{k,0}, I_{k,1}, ...)$, where t represents time. The disparity value at point P in frame t in the reference view is denoted as d_t. 3D scene flow at point P is denoted as (u, v, w), where u, v are actually the components of optical flow vector. w is defined as the disparity motion $d_{t+1} - d_t$.

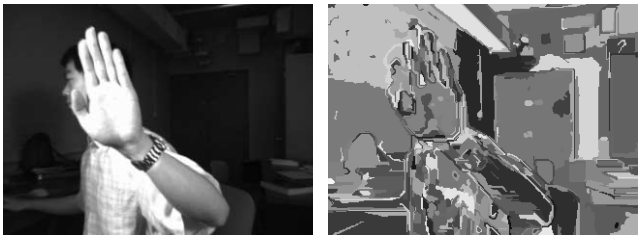
To compute 3D scene flow and structure from multiview image sequences, a natural thought is to compute optical flow (u, v) and disparity d_t separately in the reference image sequence. Then 3D scene flow is as simple as $(u, v, d_{t+1} - d_t)$. Ideally, if the optical flow and the disparity are accurate enough, this is correct. However, in practice both optical flow and disparity analyses are subject to their own inherent difficulties and they are still research topics in their own rights. Furthermore, it is easy to notice that optical flow computation only utilizes the constraints from one camera, and stereo matching ignores the temporal information. It is desirable that both spatial and temporal information from all different views contributes to 3D motion and structure estimation.

After estimating 3D scene flow, we can easily get optical flow from different views by projecting 3D scene flow onto the corresponding image plane. If the corresponding 3D object point is not occluded in the new view, we expect that the projected optical flow is more accurate since more information is employed during scene flow estimation. Also, scene flow can be used to guide the stereo matching in the next frame, thus improving the efficiency.

2.1 Image Segmentation

For the experiments described in this paper we have used the graph-based image segmentation proposed in [9]. As discussed before, we assume that there is no large motion or disparity discontinuities within an image segment so that smoothness constraint can be applied to each image segment. This guarantees the smoothness in textureless regions because a textureless region tends to be grouped as one segment. On the contrary, we do not enforce smoothness across the boundaries for actually smooth but highly textured regions because these regions are easily over-segmented. However, this is usually not a problem since motion and structure estimation tends to be reliable in textured regions even without the smoothness constraint. In other words,

ISBN 0-7695-1272-0/01 $10.00 (C) 2001 IEEE



(a) (b)

Figure 1: (a): Input image. (b): Segmentation result.

over-segmentation can be tolerated to some extent in our framework. Since smoothness is not applied to the entire image, sharp motion/depth boundaries can be maintained. A typical result of this image segmentation algorithm is shown in Figure 1.

2.2 Initial Stereo Matching

To get reliable 3D scene flow, we require that the initial disparity map be smooth and detailed. We also hope that continuous and even surfaces produce a region of smooth disparity values with their boundary precisely delineated, while small surface elements are detected as separate distinguishable regions. Furthermore, an ideal initial stereo matching algorithm should explicitly identify and report the occluded area and provide a confidence measurement of the computed disparity at every image point. As will be discussed later, the occlusion and confidence information is important in 3D scene flow estimation. Though obviously desirable, it is not easy for a stereo algorithm to satisfy all these requirements at the same time.

Inspired by the work of cross-validation [10], image segmentation, and prediction error testing [18], we propose a hierarchical rule-based stereo matching algorithm. A set of rules for depth hypothesis is defined to guide the matching process. The output includes a disparity map, an occlusion map, and a confidence map.

First, a correlation volume $\mathcal{C}(x, y, d)$ is computed between image $I_{0,0}$ and $I_{1,0}$ along the epipolar line. The measured disparity is the one with the largest matching score. We perform the correlation twice by reversing the roles of the two images and consider as valid only those matches for which we measure the same depth at corresponding points when matching from $I_{0,0}$ to $I_{1,0}$ and from $I_{1,0}$ to $I_{0,0}$. Following this method, we compute the valid disparities between $I_{0,0}$ and $I_{2,0}, I_{3,0}, ..., I_{N-1,0}$, respectively. Then we merge the results together and get a sparse initial "valid" disparity map. If two sets of views produce different valid disparities, then the one with higher matching score wins. Also, the matching score in the correlation volume is updated accordingly. To further increase the density of the initial valid disparity map, a pyramid control strategy is employed as suggested in [10]. Valid matches from two resolution levels are merged. To further increase the $Signal/Noise$ ratio, we filter out those valid points that are either isolated or have very large standard deviation in a small neighborhood. Finally, we get a initial disparity map with few errors.

Second, we segment the reference image into small regions. We then label each image segment as following:

$$L(s) = \begin{cases} VALID & if \ \ r \geq \alpha_1; \\ SEMIVALID & if \ \ \alpha_2 \leq r < \alpha_1; \\ INVALID & if \ \ r < \alpha_2, \end{cases} \quad (1)$$

where r is the ratio of valid disparity points in segment s, and $VALID, SEMIVALID$ and $INVALID$ are all symbolic values. $VALID$ means that we have high confidence on the disparity map within segment s. $INVALID$ means low confidence, and $SEMIVALID$ means medium confidence. This labeling method reflects an assumption we have made: image segments where the valid disparity points are dense are more reliable. Generally, experiments have shown that this assumption holds [10].

Third, we define a set of rules that guides the stereo hypothesis process according to the different labels of the segments:

Rule 1: If $L(s)$ equals $VALID$, then the disparity in this segment is filled by interpolating segment s.

Rule 2: If $L(s)$ equals $SEMIVALID$, then search all the neighbors of segment s. If one neighbor n satisfies the following criteria:

a. $L(n)$ equals $VALID$,

b. the disparity of n is very similar to that of s,

c. the intensity of n is very similar to that of s,

then store the segment number in an array K. After checking all the neighbors, if K is not empty, find the segment k in K which has the most similar disparity compared with segment s. Then the disparity in segment s is filled by interpolating segments s and k as if they were one large segment. Set $L(s)$ to $VALID$.

Rule 3: If $L(s)$ equals $INVALID$ or $SEMIVALID$, then hypothesize the disparity in segment s by using one of the $VALID$ neighbors k. Then warp the image of this segment to other views by using the hypothesized disparity [18]. The matching score $M (0 \leq M \leq 1)$ between the warped image and original image is stored. If the highest matching score corresponding to a $VALID$ neighbor (segment k) satisfies $M > T$ (T is a positive constant), then the disparity in segment s is filled by interpolating segments s and k as if they were one large segment. Set $L(s)$ to $VALID$.

Rule 4: Same as Rule 3 except set T equals 0.

The interpolation used in our algorithm is a mem-

ISBN 0-7695-1272-0/01 $10.00 (C) 2001 IEEE

brane model. The energy function can be defined as

$$\varepsilon = \iint \beta(d-d_v)^2 + \lambda_x(\frac{\partial d}{\partial x})^2 + \lambda_y(\frac{\partial d}{\partial y})^2 dxdy, \quad (2)$$

where d_v is the valid disparity within the segment computed by cross-validation. β is the normalized correlation score if valid disparity exists at that point, otherwise β is set to 0. In Rule 1, λ_x and λ_y are two positive numbers controlling the amount of smoothness along x direction and y direction, respectively. However, from Rule 2 to Rule 4, λ_x and λ_y are defined as

$$\lambda_x = c_x L_{norm}(\frac{\partial I}{\partial x}), \quad \lambda_y = c_y L_{norm}(\frac{\partial I}{\partial y}), \quad (3)$$

where I is image intensity and c_x and c_y are small positive numbers. L_{norm} is a piecewise linear function defined as

$$L_{norm}(x) = \begin{cases} 1 & if \;\; x < x_0; \\ \frac{x_{max}-x}{x_{max}-x_0} & if \;\; x_0 \le x \le x_{max} \end{cases} \quad (4)$$

where x_{max} is the maximum value of variable x and x_0 is the median value.

The reason that we use adaptive weight from Rule 2 to Rule 4 is to further enforce the depth boundaries. In Rules 2, 3 and 4, interpolation is applied to two adjacent segments. This means the smoothness terms in membrane model are actually applied across the segment boundary. Adopting adaptive weights defined in Eq. 3 prevents over-smoothing among those two segments.

In our algorithm, Rule 1 to Rule 4 are applied sequentially on each image segment. Rules 2, 3 and 4 need to be applied iteratively until there is no more updated segment before moving to the next Rule. In Rule 2, the similarity measurement of intensity/disparity between two segments is simply the absolute difference of their average values. This is because image segmentation guarantees that the intensity within one segment is very similar. Also, if the image is not seriously under-segmented, the disparity variation within a segment should not be large. The worst case happens when very slanted surface exists in the scene. In that case disparity variation in some segments may be large, and Rule 2 tends to reject the hypothesis from the neighborhood. However, this does not matter because Rule 3 or 4 still have good chances to make correct hypothesis. In fact, Rule 2 is designed to deal with over-segmentation. Rule 3 and Rule 4 are separated because we want to give higher priority to those $INVALID$ segments with larger matching score after warping. It is to be emphasized that the sequence of applying these rules are important. The general rule is that segments with more valid information should be processed first.

It is straightforward to decide the confidence map and occlusion map from the above algorithm. The disparity confidence measurement at each point is assigned in a hierarchical manner: we first assign maximum confidence values to three different segments (*e.g.*, assign 1.0 to $VALID$ segments, 0.8 to $SEMIVALID$ segments and 0.5 to $INVALID$ segments). Each point's confidence should not exceed the maximum confidence of the segment containing this point. Within each segment, we assign the confidence measurement according to the matching score.

After we get the confidence map, the occluded areas are detected by setting a threshold on the confidence map. If the confidence measurement of a point is below the threshold, we label this point as occluded. The results of our stereo matching algorithm is shown in Figure 2.

2.3 Motion Constraints

The motion constraints we use in our algorithm actually combine optical flow constraints from every single camera. The optical flow constraint in camera i can be represented as

$$\frac{\partial I_i}{\partial x}u_i + \frac{\partial I_i}{\partial y}v_i + \frac{\partial I_i}{\partial t} = 0. \quad (5)$$

Since we suppose other cameras are all in standard set up with the reference camera, it is easy to derive a way to combine the optical flow constraints. Suppose the focal length of camera 0 is f. At frame t, a 3D object point $\mathbf{P}$ is at (X_t, Y_t, Z_t). If we use the camera coordinate of the reference camera as the world coordinate, then the projection position of point $\mathbf{P}$ on the image plane of camera C_0 at frame t is

$$x_t = \frac{X_t f}{Z_t}, \quad y_t = \frac{Y_t f}{Z_t}. \quad (6)$$

Now we use a two camera set up as an example to show how to combine optical flow constraints from different cameras. Suppose cameras C_0 and C_1 form a standard set up and have the same focal length f and the base line b is along X axis. If at frame t 3D point $\mathbf{P}$ is projected at (x_t, y_t) in the reference view, then the projection position (x'_t, y'_t) of $\mathbf{P}$ on camera C_1 is

$$x'_t = x_t + \frac{bf}{Z_t}, \quad y'_t = y_t. \quad (7)$$

At frame $t+1$, point $\mathbf{P}$ is projected at (x_t+u, y_t+v) on camera C_0, and the projection position (x'_{t+1}, y'_{t+1}) of $\mathbf{P}$ on camera C_1 is

$$x'_{t+1} = x_t + u + \frac{bf}{Z_{t+1}}, \quad y'_{t+1} = y_t + v. \quad (8)$$

Since $d_t = \frac{bf}{Z_t}$ and $w = d_{t+1} - d_t$, the optical flow (u', v') of point $\mathbf{P}$ on camera C_1 is

$$u' = u + \frac{bf}{Z_{t+1}} - \frac{bf}{Z_t} = u + w, \quad v' = v. \quad (9)$$

ISBN 0-7695-1272-0/01 $10.00 (C) 2001 IEEE

So, the combined motion constraint of camera C_0 and C_1 can be represented as

$$\begin{aligned}\mathcal{E}_m =&(\frac{\partial I_0}{\partial x}|_{x,y}u + \frac{\partial I_0}{\partial y}\ |_{x,y}\ v + \frac{\partial I_0}{\partial t}\ |_{x,y})^2 \\ &+ \kappa(\frac{\partial I_1}{\partial x}|_{x+d,y}(u+w) + \frac{\partial I_1}{\partial y}|_{x+d,y}v + \frac{\partial I_1}{\partial t}|_{x+d,y})^2\end{aligned} \tag{10}$$

where κ is the confidence measurement of disparity (obtained from the stereo matching algorithm) if $\mathbf{P}$ is visible in camera C_1, otherwise it is 0.

It is straightforward to extend this combination to situations where more than two cameras are utilized. Ideally, the motion constraints should combine all the optical flow constraints from all the cameras. However, under a multiview setup, occlusion is almost unavoidable. Only the optical flow constraints from some of the cameras are usable. Note that if the object point $\mathbf{P}$ is not visible in other cameras except the reference camera, the combined motion constraint degrades to normal optical flow constraint.

2.4 Hard Constraints

In initial stereo matching, we can easily add hard constraints by setting the weight β in Eq. 2 to a very large number for the points with high confidence measurement. This makes the disparity map more accurate. Similarly, hard constraints can be added in the temporal domain. We perform correlation twice on two consecutive frames. Suppose P is a point in the reference frame t. We first search the correspondence of point P in frame $t+1$ using correlation. The search area is within a window delimited by the possible maximum motion. We then exchange the roles of the two frames and find the correspondence again. If at the correspondence points we have the same motion measurement in both cases, we consider the 2D motion (u_h, v_h) at the image point as valid motion. The hard constraint at a point in temporal domain can be represented as

$$\mathcal{E}_h = \mu c((u-u_h)^2 + (v-v_h)^2), \tag{11}$$

where (u_h, v_h) is the valid motion found by cross-validation, and c is a large constant. μ is the normalized matching score while searching for motion correspondence if valid motion has been measured, otherwise it is 0.

2.5 Formulation 1: 3D Scene Flow Estimation

If we assume that the initial disparity map is accurate enough, we can formulate the problem as: given N image sequences captured by N different cameras and an accurate initial disparity map, compute 3D scene flow (u, v, w) at every point in the reference image. Combining the motion constraints and hard constraints, we have an energy function

$$\mathcal{E}_1 = \iint (\mathcal{E}_m + \mathcal{E}_h + \mathcal{E}_s)dxdy \tag{12}$$

where $\mathcal{E}_m$ and $\mathcal{E}_h$ are defined in previous sections. $\mathcal{E}_s$ is the smoothness term defined as

$$\begin{aligned}\mathcal{E}_s = \gamma((\frac{\partial u}{x})^2 + (\frac{\partial u}{y})^2 + (\frac{\partial v}{x})^2 + (\frac{\partial v}{y})^2 + \\ (\frac{\partial w}{x})^2 + (\frac{\partial w}{y})^2).\end{aligned} \tag{13}$$

In our algorithm, we want to minimize the above energy function within individual image segments to enforce motion boundaries. However, the recovery of segmented or piecewise smooth flow field is notably difficult [5]. Under a multiview setup, we may have more constraints than what we have by using only one camera. But we also have more unknowns (z motion) to solve. There are even more difficulties when a segment in the reference image is invisible in a lot of other cameras. In an extreme case, a segment in the reference camera is invisible in all the other cameras. This means we may have only two constraints (one optical flow constraint and one smoothness constraint) at a point in this segment in order to solve three unknowns (u, v, w). Mathematically speaking, w is not well defined by these constraints and the algorithm may converge incorrectly. Thus we need more information from the neighborhood to propagate into these points.

To deal with the occluded regions, we adopt a multiresolution strategy to minimize the energy. First, we minimize the energy function on the entire image. This forces more neighborhood information to propagate towards the occluded points. Second, we use the results from the first step as initial values to minimize energy function within each segment. Obviously, for segments that are totally occluded in all the other views, the second step is not necessary. Experiments (Figure 3) show that this strategy can still maintain reasonable motion boundaries.

As suggested in [11] and [16], an iterative relaxation method or conjugate gradient method may be used to minimize Eq. 12. In our experiments, conjugate gradient search is used. It is also clear that the energy function is very similar to gradient-based optical flow energy function. In fact, this formulation can be thought of as a natural extension of optical flow computation under a multiview setup. The difference is that all the optical flow constraints from different views contribute to the minimization wherever possible. Also, newly added hard constraints makes the algorithm more stable and accurate.

ISBN 0-7695-1272-0/01 $10.00 (C) 2001 IEEE

2.6 Formulation 2: Integrated 3D Scene Flow and Structure Estimation

The previous formulation assumes that we have already got very accurate disparity map in initial stereo matching. However, initial stereo matching may be inaccurate and noisy because stereo matching is essentially an under-constrained problem due to occlusion, lack of texture, *etc.* Furthermore, as we discussed before, stereo matching algorithm normally ignores temporal information. It is reasonable to think that by considering motion constraints, we may get better disparity map. This means we formulate the problem as computing a four dimensional vector (u, v, w, d) at every point on the reference image, where the initial disparity is used as an initial guess. However, with serious occlusion and limited number of cameras, this formulation is even more difficult because we now need to solve four unknowns at every point. We need at least four independent constraints to make the algorithm stable. This means if we use Eq. 12, at least three cameras should be used for this formulation so that we can have three optical flow constraints, one smoothness constraint, and maybe one hard constraint. Thus new constraints in addition to those used in Eq. 12 are desired.

Considering the correlation volume $\mathcal{C}(x, y, d)$ we computed from initial stereo matching, a new constraint on disparity can be established. For image point (x_p, y_p), by using planes $x = x_p$ and $y = y_p$ to carve the correlation volume, we can get a one dimension function $\mathcal{C}_{x_p,y_p}(d)$. Obviously the disparity should maximize the value of $\mathcal{C}_{x_p,y_p}(d)$. Thus another energy term, stereo constraint, can be defined as

$$\varepsilon_c = -\tau \mathcal{C}_{x,y}(d), \tag{14}$$

where τ is a positive constant if the point is not occluded in the corresponding camera; otherwise it is 0. Also, if we have high confidence measurement for initial disparity map, we can add another energy term

$$\varepsilon_i = \zeta(d - d_i)^2, \tag{15}$$

where d_i is the initial disparity and ζ is its confidence measurement.

Thus the new energy function can be defined as

$$\mathcal{E}_2 = \int\int (\mathcal{E}_m + \mathcal{E}_h + \mathcal{E}_s + \mathcal{E}_c + \mathcal{E}_i) dxdy. \tag{16}$$

Another change of energy function in this formulation is that the smoothness term ε_s should include the smoothness measurement of disparity defined as

$$(\frac{\partial d}{x})^2 + (\frac{\partial d}{y})^2. \tag{17}$$

Again, we can use the multiresolution strategy discussed in Section 2.5 to minimize the energy function in each image segment.

In both formulations, to take advantage of the state of the art in optical flow estimation, we can first compute the optical flow in the reference view, then use the results as the initial guess while minimizing the above energy functions. This makes the algorithm converge faster and be more robust.

3 Experimental Results

We implemented the proposed algorithm on a PC platform. Digiclops system (PointGrey Inc.) is used to acquire multiview image sequences. Digiclops has three calibrated progressive scan CCD cameras in standard setup. It connects with PC through Firewire (IEEE1394) and captures image sequences at a speed of about 16 frames/sec. This device provides us real-time rectified image sequences and camera calibration parameters.

We first applied our stereo matching algorithm on a snapshot of Digiclops. Figures 2 (a) and (b) show the left and right (reference) views. Top view is not shown here. The initial valid matches are illustrated in (c). For the purpose of comparison, we applied a direct method [10] (image gradient based interpolation) to the three views and the resultant disparity map is shown in (d). The second row illustrates the output of our algorithm. (e) displays the valid segments after we applied Rule 1 and 2 on the images. (f) is our final result. Compared with (d), it is clear that our algorithm maintains very good depth boundaries. The shapes of the chair, hand, arm, and the silhouette of the person, are very clear. The fattening effects existing in the direct method is also heavily reduced. As mentioned before, our algorithm not only produces dense disparity map (f), but also occlusion detection (g) and confidence measurement (h). (g) shows the regions which are visible in the right (reference) view but invisible in the left view. It is clear that our algorithm captures the occluded areas correctly. (h) is a dense confidence map corresponding to the disparity map. The brighter a point, the more confident we are in its disparity value. It is worthwhile to note that the occlusion detection in (g) is not directly related to the confidence map shown in (h). This is because (h) is based on all the three views while (g) is only based on two (left and right) views.

The image resolution is 320×240. The minimum disparity is 2 pixels and the maximum disparity is 70 pixels (almost 22% of the image). This very large disparity range means that if we only have two cameras, serious occlusion occurs (as shown in Figure 2 (g)) and we can not get any reliable information in the occluded regions. Under our three camera setup, the top view provides rich information in these occluded areas. When computing the sparse valid disparity map (Figure 2 (c)), we use a correlation window with a radius of 5. When labeling the segments according to valid point density,

ISBN 0-7695-1272-0/01 $10.00 (C) 2001 IEEE

we set $\alpha_1 = 0.9$ and $\alpha_2 = 0.5$.

We also implemented the formulations to estimate 3D scene flow/structure described in Section 2. Figure 3 illustrates our results. (a) and (b) are two consecutive reference frames. The left hand of the person is moving upper left and towards the camera. The right hand of the person is moving away from the camera. This motion is like a small fraction of the natural hand movement of a jogging person. We first applied our stereo matching algorithm to get the initial disparity map (as shown in (c)). To test the robustness of our stereo algorithm, we used exactly the same set of parameters here as what we used to compute the results in Figure 2. From the result we can see that the algorithm still maintains sharp depth boundaries. (e) and (f) are the results of applying formulation 1 to these images. (e) is the projected 2D motion of 3D scene flow on the reference view. (f) is the z (disparity) motion. Brighter means the object is moving towards the camera. From the results we can clearly see that we obtained the correct dense 3D scene flow field: the left hand is moving upper left and towards the camera, while the right hand is moving away from the camera. (d), (g) and (h) show the results produced by formulation 2. From the results, we can see few differences between formulation 1 and formulation 2. This is because the moving areas between these two frames are small and the disparities do not benefit much from the motion constraints. We expect that when large motion areas exist (*e.g.*, camera ego-motion), formulation 2 should be more appropriate.

4 Conclusion and Future Work

In this paper, we have presented a novel method to estimate 3D scene flow and structure from multiview image sequences. A hierarchical rule-based stereo matching algorithm is proposed to estimate the initial disparity map. Available constraints in a multiview setup are investigated. We combine optical flow constraints from different views, hard constraints, and stereo constraints, *etc.* to make the algorithm more accurate and robust. The basic advantage of our algorithm is the ability to exploit all the available constraints in one minimization framework. Two formulations to estimate 3D scene flow and structure under different assumptions have been proposed. All the algorithms have been implemented and applied on real imagery. Promising experimental results are shown.

Our future work includes trying to utilize hierarchical basis function during the minimization to improve the efficiency. We are also carrying out extensive experiments to further evaluate the performance of the two different formulations under different situations.

References

[1] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *IJCV*, 2(3):283–310, January 1989.

[2] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. In *CVPR92*, pages 236–242, 1992.

[3] J.L. Barron, A.D. Jepson, and J.K. Tsotsos. Determination of egomotion and environmental layout from noisy time-varying velocity in binocular image sequences. In *IJCAI87*, pages 822–825, 1987.

[4] M.J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow-fields. *CVIU*, 63(1):75–104, January 1996.

[5] M.J. Black and A.D. Jepson. Estimating optical-flow in segmented images using variable-order parametric models with local deformations. *PAMI*, 18(10):972–986, October 1996.

[6] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. In *ICCV99*, pages 377–384, 1999.

[7] L.G. Brown. A survey of image registration techniques. *Surveys*, 24(4):325–376, December 1992.

[8] U.R. Dhond and J.K. Aggarwal. Structure from stereo: A review. *SMC*, 19(6):1489–1510, November 1989.

[9] P.F. Felzenszwalb and D.P. Huttenlocher. Image segmentation using local variation. In *CVPR98*, pages 98–104, 1998.

[10] P.V. Fua. Combining stereo and monocular information to compute dense depth maps that preserve depth discontinuities. In *IJCAI91*, pages 1292–1298, 1991.

[11] B.K.P. Horn and B.G. Schunck. Determining optical flow. *AI*, 17:185–203, 1981.

[12] T.S. Huang and S.D. Blostein. Robust algorithms for motion estimation based on two sequential stereo image pairs. In *CVPR85*, pages 518–523, 1985.

[13] T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. *PAMI*, 16(9):920–932, September 1994.

[14] R. Mandelbaum, G. Kamberova, and M. Mintz. Stereo depth estimation: A confidence interval approach. In *ICCV98*, pages 503–509, 1998.

[15] R. Mandelbaum, G. Salgian, and H. Sawhney. Correlation-based estimation of ego-motion and structure from motion and stereo. In *ICCV99*, pages 544–550, 1999.

[16] R. Szeliski. Fast surface interpolation using hierarchical basis functions. *PAMI*, 12(6):513–528, June 1990.

[17] R. Szeliski and R. Zabih. An experimental comparison of stereo algorithms. In *International Workshop on Vision Algorithms*, pages 1–19, 1999.

[18] H. Tao and H.S. Sawhney. Global matching criterion and color segmentation based stereo. In *WACV00*, pages 246–253, 2000.

[19] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. In *ICCV99*, pages 722–729, 1999.

[20] A.M. Waxman and J.H. Duncan. Binocular image flows: Steps toward stereo-motion fusion. *PAMI*, 8(6):715–729, November 1986.

[21] Y. Zhang and C. Kambhamettu. Integrated 3d scene flow and structure recovery from multiview image sequences. In *CVPR00*, pages II:674–681, 2000.

[22] C.L. Zitnick and T. Kanade. A cooperative algorithm for stereo matching and occlusion detection. *PAMI*, 22(7):675–684, July 2000.

ISBN 0-7695-1272-0/01 $10.00 (C) 2001 IEEE

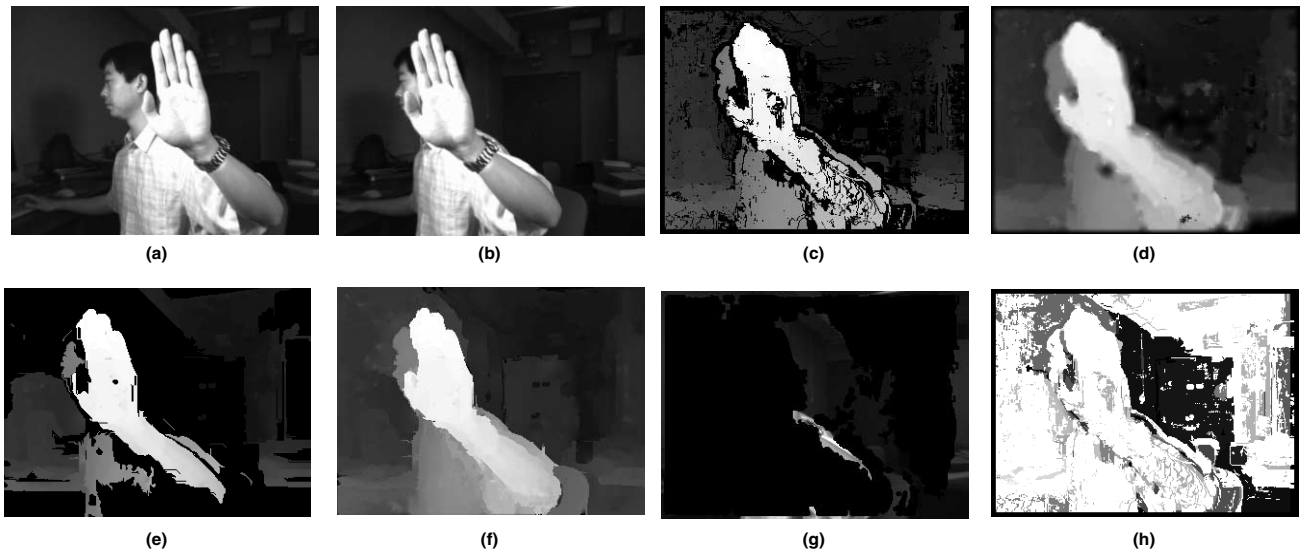


Figure 2: Stereo results on a snapshot of Digiclops system. (a) and (b) are the left and right (reference) views, respectively. Top view is not shown here. (c) is the sparse disparity map generated by merging the valid points from three views and two level Gaussian pyramid. (d) is the dense disparity map generated by a direct method. Second row shows the output of our stereo matching algorithm. (e) is the segments labeled as $VALID$ after we applied Rule 1 and 2. (f) is the final dense disparity map. (g) is the detected regions in the reference view which are occluded in the left view. (h) is the confidence measurement map of each point - brighter means higher confidence.

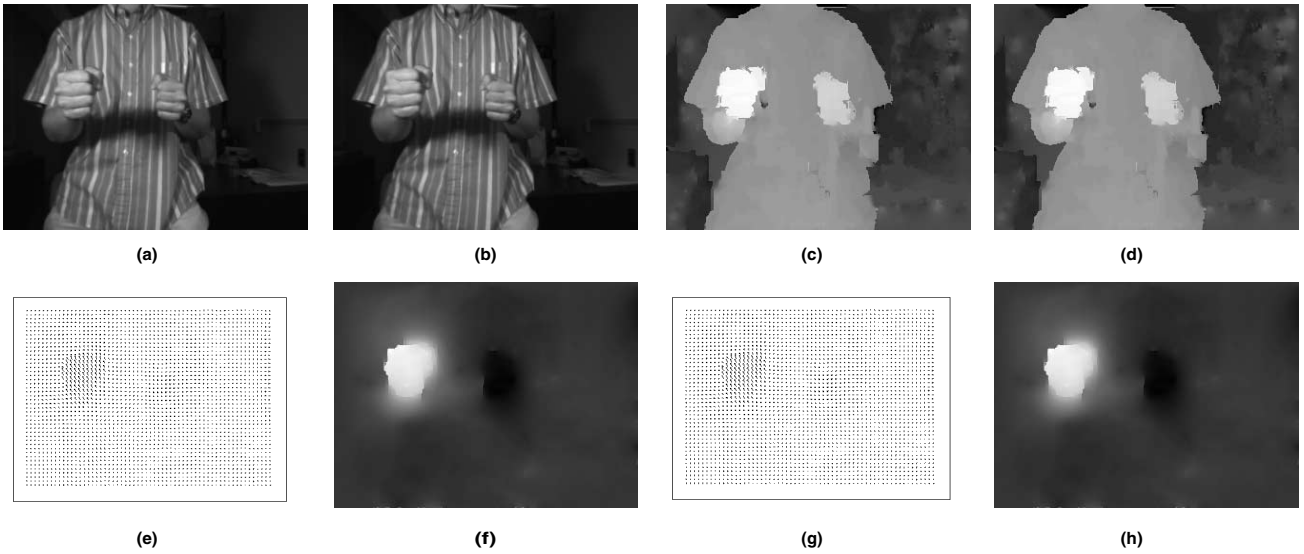


Figure 3: Scene flow and structure results. (a) and (b) are the two consecutive right (reference) views on which we need to estimate 3D scene flow and structure. Other available views are not shown here. (c) is the disparity map generated by the proposed stereo matching algorithm. (d) is the disparity map generated by formulation 2. (e) and (f) show the 3D scene flow estimated by formulation 1. (g) and (h) are by formulation 2. (e) and (g) are the projected 3D scene flow on the reference view, while (f) and (h) illustrate the z (disparity) motion.

ISBN 0-7695-1272-0/01 $10.00 (C) 2001 IEEE